

A Multi-Model, Multi-Domain Benchmark of Large Language Model Agreement with Humans and with Each Other in Sentiment Classification

Aneesh K Sajan*

Independent Researcher, Seattle, WA, USA

*Correspondence: asajannow@gmail.com; ORCID 0009-0002-5704-957X

Abstract

Large language models (LLMs) are widely used to classify sentiment, yet cross-provider comparisons on the same data are rare and inter-model agreement is largely unmeasured. We evaluate five LLMs from four providers (Claude Opus 4.7, GPT-4o, GPT-5.5, Llama 3.1 8B, and Gemini 2.5 Pro), with VADER (a rule-based lexicon) and RoBERTa (a supervised transformer) as a baseline, on five human-annotated datasets (8,872 sentences) spanning formal and informal text in financial and general domains, under one fixed zero-shot prompt. We report Macro F1 and Cohen’s kappa against human labels, agreement between every model pair, and consensus rates. Claude Opus 4.7 and GPT-5.5 lead (mean Macro F1 0.775 and 0.776; mean kappa versus humans 0.690 and 0.700) and, with GPT-4o, form a tight cluster (pairwise inter-model kappa 0.74 to 0.89); Llama 3.1 8B, which runs locally, is moderate; Gemini 2.5 Pro, served via the Batch API, scores near chance — and a controlled minimum-thinking-budget re-run rules out “overthinking” (extended deliberation harming a simple task) as the cause. All six deployed classifiers (the five LLMs and VADER) agree on only 14 to 24 percent of sentences depending on the dataset (highest on formal financial text), so model choice materially changes labels, especially on informal text. We release all predictions, code, and data.

Keywords: large language models, sentiment analysis, benchmark evaluation, inter-model agreement, zero-shot classification, generative AI evaluation

Introduction

Sentiment analysis sorts text as positive, negative, or neutral. It is one of the most widely used NLP tasks. People apply it to predict financial markets [1], manage customer experience [2], track public health [3], and monitor research [4]. Instruction-tuned large language models (LLMs) have changed how this is done. Models such as GPT-4o, Claude, and Gemini now classify sentiment well in zero-shot mode. This raises a tempting option: replace a specialist annotator or a supervised classifier with a single API call.

That option brings both promise and risk. The promise is clear. A zero-shot LLM removes the cost and delay of manual labelling and the work of training and maintaining a task-specific model. The risk is less obvious. Different LLMs are trained on different data, aligned in different ways, and built by different providers. They may disagree, in a systematic way, about the sentiment of the same sentence. When they do, the result of any later analysis depends on which model was used, and the practitioner cannot see this. A financial analyst classifying earnings-call sentiment with Claude may reach different conclusions than one using GPT-4o, and a public-health researcher tracking social media with Gemini a different signal than with Llama.

Earlier work tests LLMs on sentiment but with narrow scope, typically one provider’s models scored only against human labels [5, 6, 7] (below). To our knowledge, no study has compared all four major

providers (Anthropic, OpenAI, Meta, Google) on the same sentence-level sentiment benchmarks under one fixed protocol, and none has measured *inter-model disagreement*: how often the models agree with each other, apart from any ground truth.

This paper fills these gaps. We ask three research questions:

- **RQ1:** How accurately do LLMs from different providers classify sentiment against human gold labels?
- **RQ2:** How much do LLMs agree with each other, and which pairs disagree most?
- **RQ3:** How do accuracy and inter-model agreement change across domains (formal versus informal, financial versus general)?

We answer these through a benchmark study on five human-annotated datasets chosen by explicit inclusion criteria. Our contributions are: **(1)** a reproducible dataset-selection procedure with five inclusion criteria and an evidence table; **(2)** the first single-protocol (fixed-prompt) benchmark spanning all four major providers — Claude Opus 4.7 (Anthropic), GPT-4o and GPT-5.5 (OpenAI), Llama 3.1 8B (Meta), Gemini 2.5 Pro (Google) — on the same 2×2 domain $\times$ register grid plus a cross-domain anchor, with the deployment-relevant consequence that two practitioners using different LLMs on the same text obtain different sentiment (earlier cross-family work [33, 34] did not measure categorical agreement between models); **(3)** a full inter-model agreement analysis (pairwise Cohen’s κ for all 15 pairs, plus consensus and plurality rates); **(4)** a controlled test of the “overthinking” hypothesis — that a reasoning model’s extended deliberation degrades simple-task performance — which we reject for Gemini 2.5 Pro by re-running at the minimum thinking budget with no statistically significant change in score; **(5)** evidence that “neutral” is defined differently across datasets, bounding cross-domain comparison of neutral scores; and **(6)** calibration against a supervised transformer baseline (RoBERTa, on all five datasets), not only the VADER lexicon.

Sentiment analysis has progressed from lexicons [8, 9] through classical machine learning [10] and transformer models [11] to instruction-tuned LLMs across subjective-language tasks [59]. The standard positive/negative/neutral scheme looks simple but hides annotation difficulty: the neutral class has no single definition — investor-irrelevant in financial text [12], conversational or factual in social media [13], mixed or balanced in reviews [14] — so “neutral” is not comparable across corpora. Human agreement is typically moderate-to-substantial and lowest on neutral labels [12, 15]. Zero-shot LLM classification relies on pretrained knowledge with no task-specific training, and prompt framing strongly affects the output [16, 17]; for sentiment, direct prompting beats chain-of-thought, whose longer reasoning chains lower human-label agreement — the “overthinking” effect [5], which we test directly on Gemini 2.5 Pro (see Discussion). This motivates our single short definitional prompt with no chain-of-thought.

Disagreement between raters is itself informative: annotator disagreement carries linguistic uncertainty and should not be averaged away [18, 19], and LLM outputs vary run-to-run [20]. Recent work treats model disagreement as systematic rather than noise [47], useful for locating genuinely ambiguous items [57], and as revealing structure that single-model accuracy hides [48]. We extend this to disagreement between model families, where the cause may be ambiguity, differing domain priors, or differing alignment. We quantify agreement with Cohen’s κ and Fleiss’ κ to stay comparable with the NLP agreement literature [15, 49]; other coefficients such as Krippendorff’s α may better handle non-pairwise missingness [49].

LLM evaluation on sentiment benchmarks. Most LLM sentiment studies share two limits we address: they cover one model family (usually OpenAI’s GPT series) and score only against human labels, never checking whether providers agree. Wang et al. [7] found ChatGPT strong on binary sentiment but weak on three-way neutral, consistent with our spurious-neutral analysis (Results). Zhang et al. [6] showed fine-tuned specialists still beat zero-shot LLMs in-domain, and Fatouros et al. [21] that ChatGPT can match FinBERT on financial sentiment without fine-tuning — both within one family. The

closest work, Vamvourellis and Mehta [5], tests three OpenAI models on the Financial PhraseBank and names the “overthinking” effect we test in the Discussion. The nearest cross-family studies — Buscemi and Proverbio [33] (ChatGPT/Gemini/LLaMA on multilingual 1–10 ratings) and Wu et al. [34] (nine LLMs on aspect-based sentiment) — report large cross-model differences but do not measure categorical agreement between models. Reasoning’s value is task-dependent, hurting binary sentiment while helping fine-grained emotion [45], and prompted LLMs still trail supervised models on sarcasm [51]. Financial work shows light instruction tuning helps [52], reasoning adds latency without improving accuracy [53], encoder–LLM hybrids can win [56], and open models carry sentiment signal but lose accuracy through prompting [54]; leaderboards stress reproducible comparison [55].

Dataset selection and LLM annotation. Dataset choice strongly shapes conclusions, yet few studies state formal selection criteria; scores vary widely by domain and text type [22], and although standard multi-dataset benchmarks help [13, 30], the neutral class remains non-comparable across corpora [12, 15]. On the annotation side, LLMs can rival crowd workers [23] but need task-specific validation [24], can diverge from humans on subjective items [44, 46], and scoring against a single majority-vote label discards the spread of human opinion [50]. These tensions motivate measuring both how well a model agrees with humans (Results) and whether another provider’s model would label differently (Results). Our study differs in covering four providers, using a 2×2 domain $\times$ register grid plus a cross-domain anchor with explicit inclusion criteria (Methods), and treating inter-model agreement as a primary goal rather than accuracy alone.

Results

All metrics are computed on valid predictions only (pairwise complete case; see Methods). Macro F1 is computed over three classes for D1, D3, D4, D5; over two classes (positive, negative) for D2 SST-2.

RQ1: Accuracy vs. Human Gold Labels

Table 2: Macro F1 by Model and Dataset

Model	D1 FPB	D2 SST-2	D3 TweetEval	D4 TFNS	D5 VADER	Mean
Claude Opus 4.7	0.958	0.953	0.652	0.679	0.633	0.775
GPT-4o	0.906	0.887	0.698	0.733	0.560	0.757
GPT-5.5	0.956	0.956	0.641	0.671	0.655	0.776
Gemini 2.5 Pro	0.407	0.354	0.311	0.300	0.305	0.335
Llama 3.1 8B	0.565	0.854	0.664	0.674	0.537	0.659
VADER (lexicon baseline)	0.479	0.595	0.531	0.427	0.472	0.501
RoBERTa (supervised baseline)	0.565	0.785	0.694	0.599	0.514	0.631

*Note. The **Mean** column mixes SST-2’s two-class score with four three-class scores, so read it as a rough summary rather than an exact quantity; the per-dataset columns are the primary evidence. The ranking is the same when the mean is taken over the four three-class datasets only (Claude 0.731, GPT-5.5 0.731, GPT-4o 0.724). RoBERTa is a supervised baseline fine-tuned on tweet sentiment.*

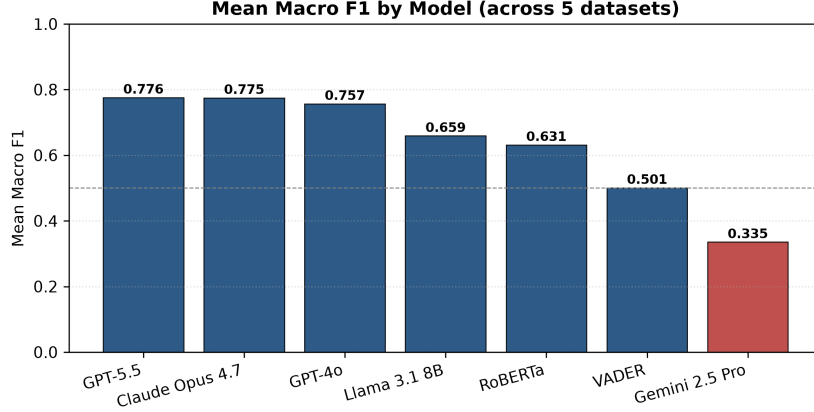


Figure 1: Mean Macro F1 by model across the five datasets. Claude Opus 4.7 and GPT-5.5 lead; Gemini 2.5 Pro falls below chance.

Figure 1 ranks the models by mean Macro F1; Supplementary Fig. S1 shows the full model $\times$ dataset breakdown.

Comparison with a supervised baseline. RoBERTa, a supervised transformer fine-tuned for tweet sentiment, calibrates the zero-shot results against a stronger reference than the VADER lexicon. Its mean Macro F1 (0.631) clearly beats VADER (0.501) but trails every frontier LLM except the near-chance Gemini 2.5 Pro, and even trails Llama (0.659). It performs best on TweetEval (0.694, comparable to GPT-4o’s 0.698); however, because this model was trained on TweetEval’s own training split, that result reflects near-in-distribution evaluation rather than clean cross-domain generalisation, and it falls to 0.514–0.599 on the financial and cross-domain datasets. The top three frontier LLMs, by contrast, remain strong across all five datasets without any task-specific training — so the comparison is not merely against a lexicon.

Table 3: Cohen’s κ (Model vs. Human) by Dataset

Model	D1 FPB	D2 SST-2	D3 TweetEval	D4 TFNS	D5 VADER	Mean
Claude Opus 4.7	0.929	0.878	0.475	0.509	0.657	0.690
GPT-4o	0.832	0.716	0.531	0.579	0.509	0.633
GPT-5.5	0.926	0.902	0.470	0.504	0.698	0.700
Gemini 2.5 Pro	0.138	−0.062	0.004	−0.002	0.028	0.021
Llama 3.1 8B	0.325	0.628	0.488	0.486	0.460	0.477
VADER (baseline)	0.274	0.247	0.311	0.133	0.334	0.260
RoBERTa (supervised baseline)	0.339	0.501	0.516	0.373	0.419	0.430

Per-model accuracy shows the same ordering as Macro F1 and is reported in Supplementary Table S4. Landis & Koch [31] interpretation: $\kappa < 0.20$ = slight; 0.21–0.40 = fair; 0.41–0.60 = moderate; 0.61–0.80 = substantial; > 0.80 = almost perfect.

Figure 2 visualises human-to-LLM agreement with Landis–Koch bands. GPT-5.5 and Claude reach “almost perfect” agreement on the formal datasets (FPB, SST-2) and “substantial” on VADER, but all models — including the leaders — drop to only “moderate” on the informal social-media datasets (TweetEval, TFNS); no model yet matches human judgement on informal text, and Gemini’s bars hover at or

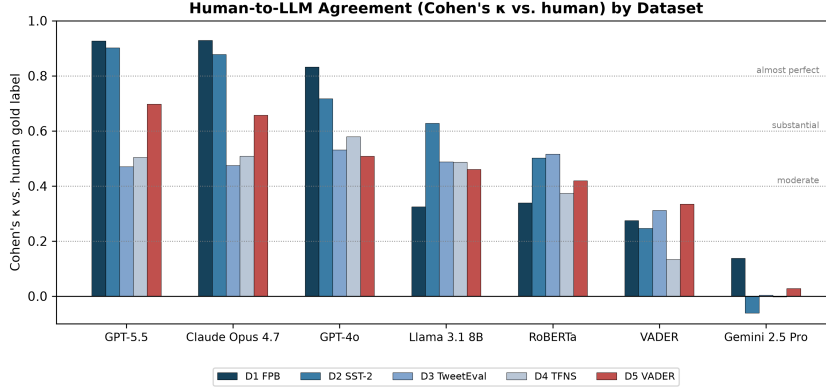


Figure 2: Human-to-LLM agreement (Cohen’s kappa versus human gold label) for each model on each dataset, with Landis-Koch interpretation bands.

below zero throughout.

Statistical significance. We confirmed the ranking with two tests. Bootstrap 95% CIs (1,000 resamples, seed = 42; Supplementary Table S5) fully overlap for Claude and GPT-5.5 ([0.767, 0.783] for both), place GPT-4o just below (upper bound 0.765), and separate the rest by wide margins. McNemar’s test on per-sentence correctness for all 15 pairs (continuity-corrected, Bonferroni $\alpha = 0.05/15 = 0.0033$) finds **14 of 15 differences significant; the sole exception is Claude vs. GPT-5.5** ($\chi^2(1) = 0.5$, $p = 0.48$) — so the two leaders are statistically indistinguishable while every other gap is real, supporting a tied top tier rather than a strict ranking.

Per-Class Analysis

Per-class F1 by model (Supplementary Table S6) shows that the Neutral class F1 is averaged over the four datasets with a gold neutral class (D1, D3, D4, D5); SST-2 has no neutral gold label and contributes only positive and negative F1, so the Macro F1 column matches the per-model means in Table 2.

The spurious neutral rate on the strictly binary SST-2 (Supplementary Table S7) exposes a systematic bias: Gemini 2.5 Pro has the highest rate (32.2%), followed by Llama 3.1 8B (18.6%) and VADER (18.1%), while Claude Opus 4.7 (3.4%) and GPT-5.5 (1.1%) largely avoid it, correctly predicting polar labels for polar texts.

RQ2: Inter-Model Agreement

Table 4: Pairwise Cohen’s κ (averaged across 5 datasets)

Model	Claude	GPT-4o	GPT-5.5	Gemini 2.5 Pro	Llama	VADER	RoBERTa
Claude	—	0.760	0.889	0.030	0.601	0.292	0.478
GPT-4o	0.760	—	0.736	0.029	0.666	0.283	0.535
GPT-5.5	0.889	0.736	—	0.032	0.585	0.317	0.474
Gemini 2.5 Pro	0.030	0.029	0.032	—	0.002	0.014	0.016
Llama	0.601	0.666	0.585	0.002	—	0.242	0.518
VADER	0.292	0.283	0.317	0.014	0.242	—	0.288
RoBERTa	0.478	0.535	0.474	0.016	0.518	0.288	—

Note. RoBERTa (supervised baseline) is shown for comparison. The consensus and plurality rates below,

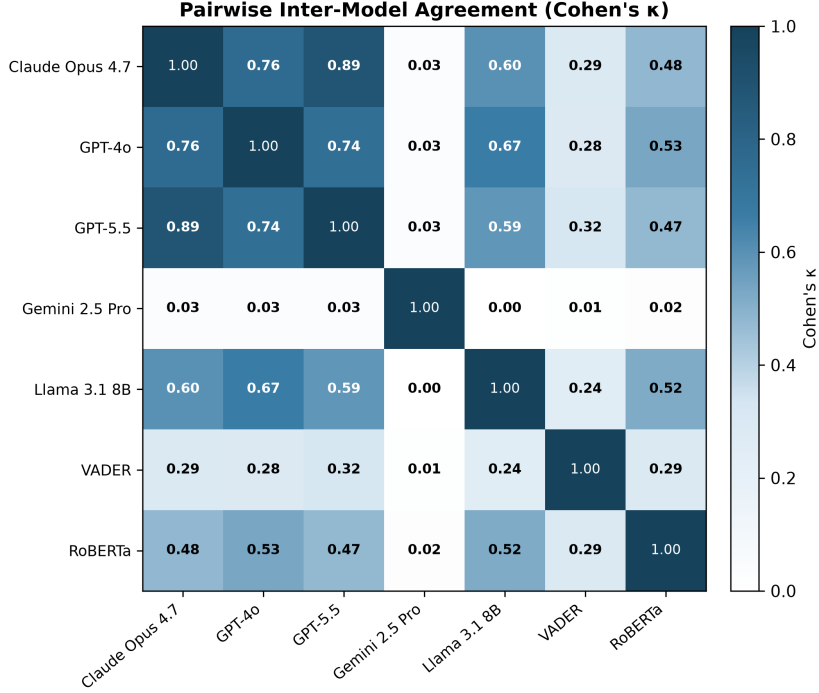


Figure 3: Heatmap of pairwise Cohen’s kappa between all model pairs, averaged across the five datasets.

Fleiss’ κ , and the McNemar significance tests are computed over the six deployed classifiers (the five LLMs and VADER); RoBERTa was added as a supervised accuracy baseline and is not part of that deployment-oriented set.

Figure 3 visualises this agreement matrix. It makes the three clusters immediately visible: the proprietary high agreement core, the moderately agreeing Llama, and the near isolated Gemini and VADER. The supervised RoBERTa baseline sits at Llama’s level of agreement with the LLMs ($\kappa = 0.47\text{--}0.54$), near zero with Gemini (0.016), and only fair with VADER (0.288) — that is, it behaves like a moderately-agreeing outsider rather than joining the proprietary core. Bootstrap 95% CIs for every pairwise κ are in Supplementary Table S18; Claude $\times$ GPT-5.5 (0.889 [0.881, 0.898]) does not overlap Claude $\times$ GPT-4o (0.760 [0.747, 0.773]), so the leading pair is statistically separable from the next-closest.

Intra-model agreement baseline. To check that measured disagreement is not merely run-to-run stochasticity, we re-ran the highest-agreeing pair (Claude and GPT-5.5) twice each on a stratified 500-sentence subset under identical settings and computed each model’s run-to-run κ . Both leaders are effectively deterministic (self-agreement $\kappa = 0.985$ and 0.994 ; $\leq 1\%$ of labels flip between identical runs), so their mutual $\kappa = 0.889$ — below each model’s intra-model ceiling — reflects genuine, systematic differences rather than noise; the argument applies *a fortiori* to the lower-agreeing pairs in Table 4. (Twenty GPT-5.5 responses exceeded the token budget on long inputs, hence $n = 480$ for that model.)

Fleiss’ κ (all 6 systems, averaged across datasets) is about 0.22 (fair). Gemini 2.5 Pro’s near random agreement with the group drives it down substantially. It is computed on the multi-model valid subsets above.

RQ3: Domain and Register Effects

Macro F1 broken down by domain/register category (Supplementary Table S11) shows a clear pattern. The two leading models achieve their highest Macro F1 on the two formal datasets, D1 (formal financial) and D2 (formal general/SST-2), and degrade on the three informal datasets. Register and domain

difficulty, not text length, dominate performance. D4 (informal financial) shows a different pattern from D1. GPT-4o outperforms Claude on informal financial text, which suggests register sensitivity. Llama’s strong SST-2 result (0.854) reflects the binary task’s relative ease. But its lower scores elsewhere keep its mean well below the proprietary models.

Qualitative Error Analysis: What Disagreement Looks Like

The aggregate metrics show *how much* models disagree; this section shows *what* they disagree about. Across the pooled 8,872 sentences, disagreement is not uniform but concentrates on identifiable linguistic phenomena. Among the four strong models (Claude, GPT-4o, GPT-5.5, Llama), 986 sentences are missed by all four, and on 865 sentences all four agree with each other yet differ from the human gold label — a signal that the disputed item may be the label, not the models. Claude and GPT-5.5, the closest pair, still disagree on 626 sentences. Table 5 gives representative cases across the categories of interest: all-models-fail, leader disagreement, ambiguous gold, and disagreement involving neutrality, sarcasm, negation, and tickers.

Table 5: Representative Disagreement Cases

Sentence (excerpt)	Gold	Model predictions	Phenomenon
“hilariously inept and ridiculous.”	positive	all four LLMs → negative	Ambiguous/questionable gold; surface-negative wording, model label defensible
“reign of fire looks as if it was made without much thought — and is best watched that way.”	positive	Claude/GPT-5.5 → negative	Sarcasm / backhanded phrasing
“or doing last year’s taxes with your ex-wife.”	negative	Claude → negative; GPT-5.5 → neutral	Leader disagreement; implicit negative simile
“you don’t have to know about music to appreciate the film’s easygoing blend...”	positive	Claude/GPT-5.5 → positive; Llama → neutral	Negation (“don’t”) trips the weaker model
“\$GM – GM loses a bull ...”	negative	Claude/GPT-5.5 → negative	Ticker + finance idiom (“loses a bull”)
“Featured #Medical #Marijuana #Stock: ... (OTC: \$PSIQ) ...”	neutral	Claude/GPT-5.5 → neutral	Promotional/spam ticker; models agree neutral

Two patterns stand out. First, a substantial share of apparent “errors” are cases where the models agree on a defensible label that differs from the gold — sarcasm and mixed-polarity movie snippets in SST-2 are the largest source — so raw accuracy understates model quality on genuinely subjective items. Second, the disagreements that remain among the strong models cluster on negation, implicit comparison, and the neutral boundary, exactly where human annotators also disagree most. Disagreement is therefore largely explainable rather than random, which reinforces the practical recommendation (see Discussion) to route borderline items to human review. An extended version of this table, with additional cases and the full model prediction set, is provided as Supplementary Table S15.

Discussion

RQ1: Which Model Best Aligns with Human Labels. Claude Opus 4.7 and GPT-5.5 are essentially tied as the strongest performers (mean Macro F1 = 0.775 and 0.776; mean κ vs. human = 0.690 and 0.700),

both reaching “substantial” agreement on average and near-perfect agreement on FPB ($\kappa \approx 0.93$) and SST-2 ($\kappa \approx 0.88$ – 0.90); GPT-4o is marginally below ($F1 = 0.757$, $\kappa = 0.633$) — substantial on average, though only moderate on TweetEval ($\kappa = 0.531$), TFNS (0.579), and VADER (0.509). The largest gap is on Financial PhraseBank, which demands the investor perspective (“Cargo traffic fell 1%” is negative for a stock even though the volume is large): Claude and GPT-5.5 exceed 0.95 there while GPT-4o reaches 0.906.

Llama 3.1 8B ($F1 = 0.659$, $\kappa = 0.477$) is “moderate” — well below the proprietary top three but above VADER — and approaches the proprietary models on informal social media (TweetEval 0.664, TFNS 0.674), a notable result for a free, locally run model; its weakest points are the VADER set ($F1 = 0.537$) and FPB (0.565). VADER reaches “fair” agreement ($\kappa = 0.260$), most competitive on social media, reaching about 76% of the best LLM’s Macro F1 on TweetEval (0.531 vs. GPT-4o’s 0.698). Gemini 2.5 Pro ($F1 = 0.335$, $\kappa = 0.021$) is an outlier, discussed below.

RQ2: Inter-Model Agreement Landscape. The inter-model agreement matrix (Table 4) reveals three distinct clusters:

Cluster 1 — high-agreement proprietary models (Claude, GPT-4o, GPT-5.5): pairwise $\kappa = 0.736$ – 0.889 , with Claude $\times$ GPT-5.5 near-perfect (0.889). Despite different providers, these models have converged on very similar sentiment representations: Claude and GPT-5.5 assign the same label to all but about 7% of sentences (they disagree on 626 of 8,872; $\kappa = 0.889$), while Claude and GPT-4o, though still highly consistent ($\kappa = 0.760$), diverge somewhat more often. This convergence may reflect shared alignment (all are RLHF-tuned on overlapping human preferences) as much as shared pretraining; disentangling the two is beyond our observational design.

Cluster 2, moderate agreement (Llama 3.1 8B): Llama agrees moderately-to-substantially with the top-3 models ($\kappa = 0.58$ – 0.67), considerably higher than its agreement with either Gemini 2.5 Pro or VADER. So despite being a smaller, open-source model, Llama has acquired similar high level sentiment representations to the proprietary models, though with more noise.

Cluster 3, low agreement (VADER, Gemini 2.5 Pro): VADER agrees only fairly with LLMs ($\kappa = 0.24$ – 0.32), as expected given its rule-based design. Gemini 2.5 Pro agrees near zero with all models ($\kappa \approx 0.00$ – 0.03), consistent with its near random performance.

Full consensus (all 6 models) is achieved on only 14–24% of sentences across datasets (Supplementary Table S9; Supplementary Fig. S2). The highest rate is on FPB (24.3%), reflecting the clearer investor perspective framing. Plurality agreement ($\geq 5/6$ models) ranges from 53–59%. So 41–47% of sentences do not even have a clear five model majority. This highlights the practical significance of model choice for borderline cases.

Disagreement is not uniform across classes: neutral predictions show the lowest inter-model agreement, consistent with neutral’s definitional heterogeneity across datasets (Methods) and prior agreement literature [12, 15].

RQ3: Domain and Register Effects. Two clear patterns emerge from Supplementary Table S11:

Domain matters more than register. The leading models achieve their highest F1 on D1 (formal financial) and show substantial degradation on D3 and D4 (informal general and financial). This suggests that financial domain knowledge is the dominant factor, not text length or formality as such. (Llama and the baselines instead peak on the binary SST-2.) VADER is most competitive relative to the LLMs on D3, reflecting its hand-crafted social-media lexicon.

Register effects within the financial domain. Comparing D1 (formal, expert annotated) and D4 (informal, crowdsourced), GPT-4o’s relative standing improves on D4 ($F1 = 0.733$ vs. Claude’s 0.679), while Claude’s absolute performance drops more sharply (from 0.958 to 0.679). This suggests Claude Opus 4.7 is more optimised for formal financial text, while GPT-4o handles the informal analyst-headline style better.

SST-2 as a calibration diagnostic. The spurious neutral rates (Supplementary Table S7) reveal a systematic bias. Models trained or aligned to expect three sentiment classes tend to over-generate neutral labels even when the gold standard is strictly binary. Gemini 2.5 Pro (32.2%) and Llama (18.6%) show the strongest such bias; VADER (18.1%) over-generates neutral for a different reason — its rule-based ± 0.05 threshold rather than a learned three-class prior. Claude Opus 4.7 (3.4%) and GPT-5.5 (1.1%) are well calibrated. Applications requiring binary (polar) classification should prompt for binary output to avoid neutral over-generation.

Gemini 2.5 Pro: A Controlled Test of the Overthinking Hypothesis. Gemini 2.5 Pro’s near-random performance (mean F1 = 0.335, mean κ = 0.021) is the most surprising result. On clearly negative FPB sentences it systematically labels falling metrics as positive — e.g., “Operating profit fell from EUR 7.9 mn to EUR 5.1 mn” → **positive**; “Cargo traffic fell 1% year-on-year” → **positive** — apparently attending to the absolute quantity rather than the directional signal (“fell”, “decreased”).

Because Gemini is reasoning-capable, a natural explanation is the “**overthinking**” effect: longer reasoning chains reduce alignment on simple tasks. This is documented for o1-like models, which spend vastly more tokens on trivial inputs with little benefit [35, 39], and Huang and Wang [45] report it degrades binary and low-class sentiment (by up to 19.9 F1 points) while helping only fine-grained emotion. Our main run used the Batch API with `max_output_tokens=2000` (~497 thinking tokens per call), making overthinking plausible.

We tested this directly. Re-running Gemini on a stratified 500-sentence subset (100 per dataset, seed = 42) via the *same* Batch API but at the minimum thinking budget (128 tokens) — the only manipulated variable — did not recover performance: mean Macro F1 = 0.328 and κ vs. human = 0.063 (95% CI [−0.006, 0.123]), statistically indistinguishable from the default-budget run on the same subset (F1 = 0.319, κ = 0.050; the full-data Gemini means are 0.335 / 0.021, computed over all sentences). We therefore reject the thinking budget as the cause. An important confound remains, however: Gemini was the *only* model evaluated through the Batch API — all others used live APIs or ran locally — so its near-chance result may reflect the Batch-API serving pathway as much as the model itself, and the minimum-budget re-run cannot break this confound because both runs use the Batch API. Two further caveats: Gemini cannot disable thinking entirely (128 tokens is the floor), and the Batch API exposes no temperature control. We therefore scope the claim narrowly: **Gemini 2.5 Pro, as served by the Batch API under default thinking, performs near chance on this task** — consistent with independently reported erratic Gemini behaviour on multilingual sentiment [33] — but we do not claim a general, model-level deficiency, which would require a live-API replication with raw outputs retained. Practitioners should validate any model *and its serving pathway* on their target task rather than assume parity. Three diagnostics rule out a parsing or dropped-output artefact (though not the pathway confound above): Gemini produced zero unresolved outputs (the lowest of any model), its confusion matrices show a coherent bias toward the class prior rather than random error, and normalisation was identical across providers (Supplementary Tables S16–S17, Note S1). Raw pre-normalisation responses were not retained, so verbatim outputs cannot be shown.

Practical Implications. For practitioners selecting models for automated sentiment annotation, the results yield four actionable recommendations.

1. **For highest accuracy on formal financial text (e.g., earnings calls, analyst reports):** Claude Opus 4.7 or GPT-5.5 are the preferred choices (F1 > 0.95 on FPB). Their near-perfect pairwise agreement (κ = 0.889) means results will be highly consistent across both.
2. **For informal financial text (e.g., analyst tweets, financial social media):** GPT-4o shows the strongest performance on D4 TFNS (F1 = 0.733) and informal domains generally. It handles the abbreviated headline style of financial Twitter more effectively than Claude on this specific register.
3. **For deployments that must run on-premise (e.g., for privacy or data-control reasons):** Llama

3.1 8B is open and runs locally. It reaches $F1 = 0.659$ on average, about 85% of the top model’s performance. For informal and social media text, Llama’s advantage is even larger relative to its mean.

4. **For reasoning-capable models:** Do not assume parity with non-reasoning peers, and do not assume a misconfiguration is to blame for poor results. Our controlled test (above) shows Gemini 2.5 Pro, as served via the Batch API, scores near chance here regardless of thinking budget. Reasoning-capable models should be validated on the specific target task before deployment. A poor result should not be attributed to extended thinking without testing that hypothesis directly.

The 41–47% of sentences without five-model plurality (Supplementary Table S9) represent genuine classification uncertainty. In high-stakes settings (e.g., financial trading signals, clinical monitoring), these borderline items should be routed to ensemble voting or human review — aggregating LLM outputs is more stable and accurate than any single model, via repeated runs [36], cross-model majority/quorum voting [42], or provider fusion that can drive error below 1% [43]. In regulated pipelines (e.g., MiFID II analyst-recommendation surveillance, complaint triage) they form a clear, auditable boundary for human review.

A deployment decision rule. The recommendations above can be applied as a short decision procedure (Supplementary Table S12).

Limitations. Internal Validity. Sampling. We used stratified random samples of up to 2,000 sentences per dataset ($n = 8,872$) rather than full datasets, preserving label proportions and length/lexical distributions (Supplementary Table S14); bootstrap 95% CIs are about ± 2 Macro F1 points.

Prompt design. A single validated definitional prompt [5] was used throughout; other prompt designs may give different results (a controlled probe of prompt sensitivity ships with the reproducibility package).

Non-determinism. Claude Opus 4.7 and GPT-5.5 do not support temperature = 0; GPT-4o and Llama use temperature = 0; Gemini uses Batch API defaults. Even nominal temperature = 0 is not strictly deterministic [37, 58], and small models vary more [38]. We bounded this by re-running the two leaders on a 500-sentence subset: intra-model $\kappa = 0.985$ and 0.994 ($\leq 1\%$ of labels flipped; Supplementary Table S10) — far below the inter-model disagreement it could be confused with. Rankings rest on large, statistically significant gaps (Supplementary Tables S5, S13). Model version strings (Table 1) were those served in June 2026.

Gemini Batch API (pathway confound). Gemini was the only model evaluated via the Batch API (all others used live APIs or ran locally), which allows no temperature control and no thinking-off condition. The minimum-thinking-budget re-run leaves performance near chance, so the result is not an artefact of extended thinking — but because the API pathway itself is uncontrolled, we cannot separate a genuine model deficiency from a Batch-API serving effect. All Gemini claims are therefore scoped to this configuration; a live-API replication with raw outputs retained is needed before generalising them.

Data contamination. All datasets are public and likely in pretraining, but benchmark items can be forgotten once training scales past the Chinchilla-optimal point [32], making verbatim memorisation of specific items unlikely. Our primary contribution (inter-model agreement) is threatened only by *differential* contamination; uniform memorisation would simply inflate agreement across the board. The differentiated structure we observe (tight proprietary cluster, moderate Llama, isolated Gemini) rules out uniform memorisation, but does not by itself exclude differential contamination, which could produce a similar pattern — so this argument bounds rather than eliminates the threat, and the cluster structure is at least equally consistent with shared alignment and differing domain priors.

External Validity. Our systems span four providers plus a lexicon and a supervised transformer baseline — broader than prior sentiment benchmarks but not exhaustive: larger open models (Llama 70B/405B), other families (Mistral, Qwen, Cohere), and further proprietary models are excluded, so cross-family

claims cover the four dominant providers as of mid-2026. Results characterise the models as accessed in June 2026 and may drift, though relative rankings and agreement structure should be more stable. All datasets are English, so findings may not transfer to other languages. Few-shot prompting and fine-tuning are excluded by design; they could narrow the gaps (especially Llama’s) and change agreement. The Gemini result reflects a specific Batch-API-plus-thinking configuration and is not its general ceiling.

Construct Validity. VADER conversion. The ± 0.05 threshold yields a tiny neutral class (2.0% of D5), making D5 behave almost as a binary task; we therefore draw neutral conclusions only from the genuinely three-class D1, D3, D4.

Mixed class counts. The mean Macro F1 averages a two-class (SST-2) with four three-class scores; it is a convenience summary, not a strictly commensurable quantity, so the per-dataset values (Table 2) and significance tests (Supplementary Table S5) are the primary evidence, and the ranking is unchanged over the four three-class datasets alone.

Neutral heterogeneity. “Neutral” differs across datasets (investor-irrelevant, conversational, or ambivalent), so neutral F1 is not compared across corpora.

Agreement metric. κ is sensitive to marginal skew (the “kappa paradox”), so we report raw consensus and plurality rates (Supplementary Table S9) alongside κ and interpret κ within, not across, datasets [49].

Conclusion. We compared five LLMs from four providers, a lexicon baseline, and a supervised transformer baseline (RoBERTa) on five human-annotated datasets spanning formal and informal registers and financial and general domains, under a single zero-shot protocol with explicit dataset-selection criteria and statistically grounded analysis (bootstrap CIs, McNemar tests, and an intra-model κ baseline).

RQ1 (accuracy): Claude Opus 4.7 and GPT-5.5 lead (mean Macro F1 ≈ 0.78 , mean κ vs. human ≈ 0.70), near-perfect on formal financial text; GPT-4o is close behind; Llama 3.1 8B is moderate (F1 = 0.66) but runs locally; VADER stays competitive on social media; and Gemini 2.5 Pro is near chance — a deficit that a controlled minimum-thinking-budget re-run shows is not driven by the thinking budget.

RQ2 (inter-model agreement): Claude, GPT-4o, and GPT-5.5 form a high-agreement cluster (pairwise $\kappa = 0.74\text{--}0.89$); an intra-model baseline ($\kappa = 0.985$ and 0.994) confirms the residual disagreement is genuine, not run-to-run noise. Only 14–24% of sentences are labelled unanimously by all six models and 41–47% lack a five-model majority, so model choice materially affects downstream results on borderline text.

RQ3 (domain effects): The frontier LLMs peak on formal financial text and degrade on informal domains (the weaker and baseline systems peak on SST-2), with GPT-4o relatively stronger on informal financial text; spurious neutral generation is a systematic bias in Gemini (32.2%), Llama (18.6%), and VADER (18.1%) that matters for binary tasks.

Future work will extend the framework to few-shot and fine-tuned settings, multilingual data, a live-API Gemini evaluation, and longitudinal tracking of how model updates change predictions on fixed benchmarks over time.

Methods

Dataset Selection Framework

Inclusion Criteria. We apply five explicit inclusion criteria (IC). These keep the dataset selection reproducible and methodologically defensible.

- **IC1 (Human annotation):** The dataset must use genuine human annotation. Expert or crowd-sourced raters read each text and assigned a sentiment label. We exclude datasets where labels come from star ratings, emoticons, hashtags, or other implicit signals.

- **IC2 (Public access):** Freely available without institutional registration, fee, or restricted access agreement.
- **IC3 (Minimum size):** At least 500 instances.
- **IC4 (Label compatibility):** Binary (2-class) or three-class labels directly convertible to {positive, negative, neutral}.
- **IC5 (Citation evidence):** Cited in at least three peer-reviewed sentiment analysis publications from 2023–2026.

Rationale for IC1 (Annotation Quality). IC1 needs explicit justification because it excludes several widely used datasets. The IMDb Large Movie Review dataset [25], Yelp Polarity [26], Amazon Reviews [27], and Sentiment140 [28] all derive labels from implicit signals, not explicit human judgment, and SemEval-2014 aspect-based sentiment [29] is excluded for label incompatibility (IC4). This study measures the *agreement between LLMs and human judgments*. So the gold standard must consist of genuine human judgments, not inferred preferences.

Evidence Table. The full evidence table applying all five criteria to every candidate dataset, both included and excluded, is provided as Supplementary Table S1.

Datasets

Five datasets pass all inclusion criteria. Together they span a 2×2 domain/register matrix (financial vs. general $\times$ formal vs. informal) plus a cross-domain anchor (Supplementary Fig. S3).

For datasets with more than 2,000 sentences, we drew a stratified random sample (seed = 42). This preserves the original class distribution. SST-2 is used in full, as its evaluation split is 872 sentences. The sampling keeps label proportional coverage for all downstream analyses. Per-dataset sample sizes, class proportions, and word-length distributions are summarised in Supplementary Table S2.

Why $n = 2,000$. The cap is driven by the financial cost of frontier-model API inference: evaluating the four API-served models over these datasets required approximately 35,000 frontier-model API calls (four models $\times$ 8,872 sentences; Llama and VADER ran locally), and $n = 2,000$ keeps per-dataset sampling error small (bootstrap 95% CI $\approx \pm 2$ Macro F1 points) while bounding that cost. **Representativeness.** Because subsampling could in principle distort the text distribution, we verified that each sample matches its full source pool not only on label proportions (preserved by stratification) but also on text length: mean word length agrees to within 0.1 words on every dataset (D1 22.3 vs. 22.4; D3 14.8 vs. 14.9; D4 12.3 vs. 12.2; D5 17.2 vs. 17.3), with matching medians and standard deviations, and matching mean character length (e.g., D3 90.9 vs. 91.3). SST-2 is the complete validation split, so no sampling error applies. Full sample-versus-population statistics are given in Supplementary Table S14.

Dataset details. Supplementary Table S3 summarises the provenance, sampling, and key modelling challenge of each dataset.

Cross-Dataset Comparison. Neutral class variation: The neutral share ranges from 0% (D2 SST-2) to 65.9% (D4 TFNS). The definitions also differ in kind across datasets. So neutral F1 is not comparable across datasets without accounting for this definitional variance.

Methodology

Models and baselines. Table 1 lists the evaluated systems and baselines.

Table 1: Models Evaluated

Model	Provider	Type	Version / API	Temp
Claude Opus 4.7	Anthropic	Proprietary	claude-opus-4-7	not supported
GPT-4o	OpenAI	Proprietary	gpt-4o	0
GPT-5.5	OpenAI	Proprietary	gpt-5.5	not supported
Gemini 2.5 Pro	Google	Proprietary	gemini-2.5-pro (Batch API)	not available
Llama 3.1 8B	Meta	Open-source	llama3.1 via Ollama	0
VADER (baseline)	Rule-based	Lexicon	vaderSentiment 3.3.2	N/A
RoBERTa (baseline)	Supervised	Transformer (tweet-trained)	cardiffnlp/twitter- roberta-base- sentiment-latest	N/A

All LLMs are evaluated in zero-shot mode: no in-context examples, no fine-tuning, and no dataset-specific prompt tuning. Temperature is set to the minimum supported value for each model to maximise output determinism. Claude Opus 4.7 and GPT-5.5 do not accept a temperature parameter and run at their default sampling setting.

Models as deployed systems. We evaluate each model as a *deployed system accessed through its provider’s standard API* under fixed decoding settings, not as an isolated set of weights. Reported differences therefore reflect the model together with its provider’s serving stack (safety filtering, default sampling, and any batch-vs-live behaviour) — which is the configuration a practitioner actually encounters. Provider, model size, alignment, and access mode co-vary across the systems compared, so results should be read at the level of deployed systems rather than as controlled ablations of any single factor.

Supervised baseline. To place the zero-shot LLMs against a stronger reference than a lexicon, we add one supervised transformer fine-tuned on in-domain sentiment data: RoBERTa [60] (cardiffnlp/twitter-roberta-base-sentiment-latest, a general 3-class model trained on tweets), run on all five datasets. It has a task-specific training advantage by design; we include it to calibrate the zero-shot results against established supervised practice (Results). It was run locally with the same label space and scoring rule as the main experiment.

Gemini 2.5 Pro, Batch API: The live API has rate limits (1,000 RPD on the free tier), so we evaluated Gemini 2.5 Pro via Google’s Batch API, which bypasses the RPM/RPD limits. The Batch API does not support a temperature parameter. Requests were submitted with `max_output_tokens=2000` to accommodate the model’s extended thinking (~497 thinking tokens per call). As detailed in the Discussion, we ran a controlled follow-up (same Batch API, minimum thinking budget). It shows the thinking budget does *not* account for Gemini’s poor performance under this Batch-API configuration; the near-chance result persists at the minimum budget.

GPT-5.5, reasoning tokens: GPT-5.5 uses about 23 reasoning tokens per call. `max_completion_tokens` was set to 100 to accommodate reasoning plus the single-word response.

Prompt Design. We use a single standardised prompt across all datasets and models:

System:

You are a sentiment classifier. Classify the sentiment of the following text using these definitions:

positive - the text expresses optimism, approval, or good news

negative - the text expresses pessimism, criticism, or bad news

neutral - the text is factual, balanced, or does not lean either way

Reply with exactly one word: positive, negative, or neutral.

User:

Text: "{sentence}"
Sentiment:

Rationale: Explicit label definitions reduce ambiguity about class boundaries, especially for neutral, without introducing reasoning steps. Vamvourellis and Mehta [5] show that chain-of-thought prompting reduces human label alignment on sentiment tasks. Wu et al. [34] independently report that vanilla zero-shot prompts outperform chain-of-thought and self-debate strategies on multilingual sentiment. We therefore exclude chain-of-thought. Fixing a single prompt is itself a deliberate control. Prompt phrasing, including small details of formatting and wording, can shift LLM classification [41], though the size of this effect is itself debated [40]. A single prompt across all datasets and all models ensures that performance differences reflect model and domain characteristics, not prompt differences. The cost is that our results characterise this prompt, not the full prompt space (the Limitations).

Output normalisation: Raw LLM outputs are normalised to {positive, negative, neutral} in two steps: (1) exact match; (2) synonym matching (bullish→positive, bearish→negative, etc.). Outputs that still could not be resolved to a label were retried twice; any that remained unresolved were excluded per metric (pairwise complete case: for each metric, rows where the relevant model or models produced no valid label are dropped).

Evaluation Metrics. Against human gold labels we report three measures per (model, dataset) cell: **Macro F1** (averaged over the classes present in that dataset’s gold labels, two for SST-2 and three otherwise), **accuracy**, and **Cohen’s κ** . For inter-model agreement we report **pairwise Cohen’s κ** for all classifier pairs (Table 4) and **Fleiss’ κ** across the six deployed systems, together with raw **full-consensus** and **plurality** rates. We assess statistical robustness two ways: bootstrap 95% confidence intervals (1,000 resamples, seed = 42) and McNemar’s test on per-sentence correctness for all pairs with Bonferroni correction (Results).

We additionally define one task-specific diagnostic:

Spurious neutral rate (D2/SST-2 only). SST-2 is strictly binary. Its gold labels contain no neutral class, so any neutral prediction is wrong. We define the *spurious neutral rate* as the percentage of a system’s SST-2 predictions that are neutral. This isolates a model’s bias toward the dominant three-class sentiment schema even when the task is binary. Values below 5% indicate good calibration to the binary task. Values above 15% indicate strong over-generation of neutral. We report this diagnostic in Results (Supplementary Table S7) and discuss its implications for binary classification deployments in the Discussion.

Data Quality. Of 53,232 total predictions (8,872 rows $\times$ 6 models), 85 (0.16%) could not be resolved to a sentiment label after normalisation and two retries. By model these are GPT-5.5 37, Llama 3.1 8B 47, and Claude Opus 4.7 1, with GPT-4o, Gemini 2.5 Pro, and VADER producing none (per-dataset breakdown in Supplementary Table S16). Notably **Gemini 2.5 Pro produced zero unresolved outputs** — the lowest of any model — so its near-chance scores (discussed later) cannot be an artefact of dropped or silently mishandled predictions. The 85 invalid outputs are excluded per metric using pairwise complete case: for per-model accuracy, only valid rows for that model are counted; for pairwise Cohen’s κ , only rows where both models in the pair produced valid labels are included.

Experiment Scale. The full experiment comprises 53,232 predictions from the six deployed classifiers (8,872 sentences $\times$ 6 systems), plus a further 8,872 predictions from the RoBERTa baseline. All LLM inference was run via provider APIs (Claude, GPT-4o, GPT-5.5, Gemini Batch API) except Llama 3.1 8B and VADER, which were run locally. The supervised baseline (RoBERTa) was run locally via the Hugging Face transformers library.

Use of AI tools. The LLMs listed in Table 1 are the object of study and were queried only as evaluated systems under the protocol above. Separately, LLM-based tools were used to assist with copy-editing and language refinement of the manuscript; all study design, analysis, and conclusions are the author’s

own. This use is also declared after the References.

References

- [1] Kirtac, K. & Germano, G. Sentiment trading with large language models. *Finance Res. Lett.* **62**, 105227 (2024).
- [2] Nugroho, A., Adi, K. & Aryasa, K. B. Literature-driven contributions to the development of LLM-based customer insight systems. *Jurnal Sisfokom* **15**, 69–74 (2026).
- [3] Sheikh, R. et al. From fear to forecast: machine learning solutions for mass psychosis in crisis scenarios. *Cureus J. Comput. Sci.* (2026).
- [4] Daruwalla, S. et al. Consistency analysis of sentiment predictions using SSAS. Preprint at <https://arxiv.org/abs/2604.15547> (2026).
- [5] Vamvourellis, D. & Mehta, D. Reasoning or overthinking: evaluating LLMs on financial sentiment analysis. Preprint at <https://arxiv.org/abs/2506.04574> (2025).
- [6] Zhang, W. et al. Sentiment analysis in the era of large language models: a reality check. Preprint at <https://arxiv.org/abs/2305.15005> (2023).
- [7] Wang, Z. et al. Is ChatGPT a good sentiment analyzer? A preliminary study. Preprint at <https://arxiv.org/abs/2304.04339> (2023).
- [8] Hutto, C. J. & Gilbert, E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. in *Proc. Int. AAAI Conf. Web and Social Media* (2014).
- [9] Liu, B. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions* (Cambridge Univ. Press, 2015).
- [10] Pang, B., Lee, L. & Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. in *Proc. EMNLP* 79–86 (2002).
- [11] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. in *Proc. NAACL-HLT* 4171–4186 (2019).
- [12] Malo, P., Sinha, A., Korhonen, P., Wallenius, J. & Takala, P. Good debt or bad debt: detecting semantic orientations in economic texts. *J. Assoc. Inf. Sci. Technol.* **65**, 782–796 (2014).
- [13] Barbieri, F. et al. TweetEval: unified benchmark and comparative evaluation for tweet classification. in *Findings of EMNLP* 1644–1650 (2020).
- [14] Socher, R. et al. Recursive deep models for semantic compositionality over a sentiment treebank. in *Proc. EMNLP* 1631–1642 (2013).
- [15] Mohammad, S. M. Best practices in the creation and use of emotion lexicons. in *Findings of the ACL: EACL 2023* 1825–1836 (2023).
- [16] Brown, T. et al. Language models are few-shot learners. in *Proc. NeurIPS* (2020).
- [17] Liu, P. et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **55**, 1–35 (2023).
- [18] Plank, B. The ‘Problem’ of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. in *Proc. EMNLP* (2022).
- [19] Baan, J. et al. Stop measuring calibration when humans disagree. in *Proc. EMNLP* 1892–1915 (2022).

- [20] Herrera-Poyatos, D. et al. An overview of model uncertainty and variability in LLM-based sentiment analysis. Preprint at <https://arxiv.org/abs/2504.04462> (2025).
- [21] Fatouros, G., Soldatos, J., Kouroumali, K., Makridis, G. & Kyriazis, D. Transforming sentiment analysis in the financial domain with ChatGPT. *Mach. Learn. Appl.* **14**, 100508 (2023).
- [22] Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A. & Benevenuto, F. SentiBench — a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Sci.* **5**, 23 (2016).
- [23] Gilardi, F. et al. ChatGPT outperforms crowd workers for text-annotation tasks. *Proc. Natl. Acad. Sci. USA* **120**, e2305016120 (2023).
- [24] Pangakis, N., Wolken, S. & Fasching, N. Automated annotation with generative AI requires validation. Preprint at <https://arxiv.org/abs/2306.00176> (2023).
- [25] Maas, A. L. et al. Learning word vectors for sentiment analysis. in *Proc. ACL* 142–150 (2011).
- [26] Zhang, X., Zhao, J. & LeCun, Y. Character-level convolutional networks for text classification. in *Proc. NeurIPS* (2015).
- [27] Ni, J. et al. Justifying recommendations using distantly-labeled reviews. in *Proc. EMNLP* (2019).
- [28] Go, A., Bhayani, R. & Huang, L. Twitter sentiment classification using distant supervision. Tech. Rep., Stanford Univ. (2009).
- [29] Pontiki, M. et al. SemEval-2014 Task 4: aspect based sentiment analysis. in *Proc. SemEval* 27–35 (2014).
- [30] Wang, A. et al. GLUE: a multi-task benchmark for natural language understanding. in *Proc. ICLR* (2019).
- [31] Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977).
- [32] Bordt, S., Srinivas, S., Boreiko, V. & von Luxburg, U. How much can we forget about data contamination? in *Proc. ICML* (2025).
- [33] Buscemi, A. & Proverbio, D. ChatGPT vs Gemini vs LLaMA on multilingual sentiment analysis. Preprint at <https://arxiv.org/abs/2402.01715> (2024).
- [34] Wu, C. et al. Evaluating zero-shot multilingual aspect-based sentiment analysis with large language models. Preprint at <https://arxiv.org/abs/2412.12564> (2024).
- [35] Chen, X. et al. Do NOT think that much for $2+3=?$ On the overthinking of o1-like LLMs. Preprint at <https://arxiv.org/abs/2412.21187> (2024).
- [36] Niimi, J. A simple ensemble strategy for LLM inference: towards more stable text classification. Preprint at <https://arxiv.org/abs/2504.18884> (2025).
- [37] Yuan, J. et al. Understanding and mitigating numerical sources of nondeterminism in LLM inference. Preprint at <https://arxiv.org/abs/2506.09501> (2025).
- [38] Pinhanez, C., Cavalin, P., Sanctos, C., Grave, M. & Primerano, Y. The non-determinism of small LLMs: evidence of low answer consistency in repetition trials of standard multiple-choice benchmarks. Preprint at <https://arxiv.org/abs/2509.09705> (2025).
- [39] Yue, L. et al. Don’t overthink it: a survey of efficient R1-style large reasoning models. Preprint at <https://arxiv.org/abs/2508.02120> (2025).
- [40] Hua, A. et al. Flaw or artifact? Rethinking prompt sensitivity in evaluating LLMs. Preprint at <https://arxiv.org/abs/2509.01790> (2025).
- [41] Pecher, B., Spiegel, M., Belanec, R. & Cegin, J. Revisiting prompt sensitivity in LLMs for text

- classification: the role of prompt underspecification. Preprint at <https://arxiv.org/abs/2602.04297> (2026).
- [42] Kamen, A. & Kamen, Y. Majority rules: LLM ensemble is a winning approach for content categorization. Preprint at <https://arxiv.org/abs/2511.15714> (2025).
- [43] Mabokela, K. R., Schlippe, T. & Wölfel, M. Large language models for sentiment analysis to detect social challenges: a use case with South African languages. Preprint at <https://arxiv.org/abs/2511.17301> (2025).
- [44] Piot, P., Otero, D., Martín-Rodilla, P. & Parapar, J. Can LLMs evaluate what they cannot annotate? Revisiting LLM reliability in hate speech detection. Preprint at <https://arxiv.org/abs/2512.09662> (2025).
- [45] Huang, D. & Wang, Z. Task complexity matters: an empirical study of reasoning in LLMs for sentiment analysis. Preprint at <https://arxiv.org/abs/2602.24060> (2026).
- [46] de-Marcos, L., Goyanes, M. & Domínguez-Díaz, A. Wisdom of the AI crowd (AI-CROWD) for ground-truth approximation in content analysis. Preprint at <https://arxiv.org/abs/2603.06197> (2026).
- [47] Ingram, W. A., Banerjee, B. & Fox, E. A. When LLMs disagree: diagnosing relevance filtering bias and retrieval divergence in SDG search. Preprint at <https://arxiv.org/abs/2507.02139> (2025).
- [48] Najera, A., Moon, A., Srinivasan, V. & Veeraraghavan, R. When models disagree: rethinking LLM evaluation for public comment analysis. Preprint at <https://arxiv.org/abs/2605.29025> (2026).
- [49] James, J. Counting on consensus: selecting the right inter-annotator agreement metric for NLP annotation and evaluation. Preprint at <https://arxiv.org/abs/2603.06865> (2026).
- [50] Inoshita, K., Zhou, X., Kawai, A. & Yada, K. LLMs capture emotion labels, not emotion uncertainty: distributional analysis and calibration of human–LLM judgment gaps. Preprint at <https://arxiv.org/abs/2604.27345> (2026).
- [51] Zhang, Y. et al. SarcasmBench: towards evaluating large language models on sarcasm understanding. Preprint at <https://arxiv.org/abs/2408.11319> (2024).
- [52] Zhang, B., Yang, H. & Liu, X.-Y. Instruct-FinGPT: financial sentiment analysis by instruction tuning of general-purpose large language models. Preprint at <https://arxiv.org/abs/2306.12659> (2023).
- [53] Huang, D. & Wang, Z. Explainable sentiment analysis with DeepSeek-R1: performance, efficiency, and few-shot learning. Preprint at <https://arxiv.org/abs/2503.11655> (2025).
- [54] Di Palma, D. et al. LLaMAs have feelings too: unveiling sentiment and emotion representations in LLaMA models through probing. Preprint at <https://arxiv.org/abs/2505.16491> (2025).
- [55] González-Bustamante, B. TextClass Benchmark: a continuous Elo rating of LLMs in social sciences. Preprint at <https://arxiv.org/abs/2412.00539> (2024).
- [56] Beno, J. P. ELECTRA and GPT-4o: cost-effective partners for sentiment analysis. Preprint at <https://arxiv.org/abs/2501.00062> (2025).
- [57] Lu, J. et al. Aligning LLM uncertainty with human disagreement in subjectivity analysis. Preprint at <https://arxiv.org/abs/2605.10415> (2026).
- [58] Fu, T. et al. Beyond reproducibility: token probabilities expose large language model nondeterminism. Preprint at <https://arxiv.org/abs/2601.06118> (2026).
- [59] Song, C., Zhang, Y., Gao, H., Yao, B. & Zhang, P. Large language models for subjective language understanding: a survey. Preprint at <https://arxiv.org/abs/2508.07959> (2025).
- [60] Loureiro, D., Barbieri, F., Neves, L., Espinosa Anke, L. & Camacho-Collados, J. TimeLMs: diachronic language models from Twitter. in *Proc. ACL: System Demonstrations* 251–260 (2022).

Acknowledgements

The author thanks the maintainers of the five public benchmark datasets and the open-source transformers, scikit-learn, scipy, and pandas projects.

Author Contributions

A.K.S. conceived and designed the study, implemented the experiments and analysis, and wrote the manuscript. As the sole author, A.K.S. is responsible for all aspects of the work.

Data Availability Statement

All sampled datasets, model predictions, and analysis code generated in this study are available in a public GitHub repository at <https://github.com/aneeshks/Multi-Domain-Sentiment-Benchmark>. The repository is available to the editors and reviewers during peer review, and a permanent Zenodo archive with a citable DOI will be added at acceptance. Derived Financial PhraseBank subsets and their predictions are released under the same CC BY-NC-SA 3.0 terms as the source; other released subsets retain their original licences. The five source datasets are publicly available from their original repositories: Financial PhraseBank (HuggingFace `gtfintechlab/financial_phrasebank_sentences_allagree`, CC BY-NC-SA 3.0); SST-2 (HuggingFace `stanfordnlp/sst2`); TweetEval (HuggingFace `cardiffnlp/tweet_eval`, CC 3.0); Twitter Financial News Sentiment (HuggingFace `zeroshot/twitter-financial-news-sentiment`, MIT); and the VADER validation set (github.com/cjhutto/vaderSentiment, MIT).

Competing Interests

The author declares no competing interests.

Funding

The author received no specific external funding for this study.

Ethics Approval

Not applicable: this study uses only publicly available, previously published benchmark datasets and collects no new human-subjects data. The social-media datasets (TweetEval, TFNS) are redistributed as originally released and may contain user handles and cashtags; we use them under their original licences and in line with Nature’s user-generated-content policy, report only aggregate statistics, and reproduce short excerpts solely for error analysis. “De-identified” here means that no additional personal data were collected, inferred, or linked beyond the public text as distributed.

Use of Large Language Models

LLMs are the **object of study** in this work. They were queried only as evaluated systems, through their APIs, under the zero-shot protocol in Section 6. LLM-based tools were also used to copy-edit the manuscript, refine the language, and help verify references. All scientific content, methodology, analysis, and conclusions are the authors’ own.